Journal of Information Systems Engineering and Management
2025, 10(18s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Topic Discovery in the Digital Quran: A Text Mining Approach

Amro Ali Badawy [1], Ebrahim Elhinawy [1], Ahmad Salah [2,1], Mahmoud A. Mahdi [1]
[1] *Department of Computer Science, Faculty of Computers and Informatics, Zagazig University, Egypt*
[2] *College of Computing and Information Sciences, University of Technology and Applied Sciences, Ibri, Sultanate of Oman*
microtouch_group@yahoo.com; ahmad@zu.edu.eg; mahdy@ zu.edu.eg

ARTICLE INFO

Received: 18 Dec 2024

Revised: 28 Jan 2025

Accepted: 12 Feb 2025

ABSTRACT

The research addresses the thematic analysis of the Quran, using for this purpose the Surah Al-Kahf and Surah An-Naml via computational approaches. This work outlines the design of a systematic method for understanding the deeply intricate moral, ethical, and spiritual understandings in those chapters. With the help of a Latent Dirichlet Allocation (LDA)-a type of topic modeling algorithm-the current study will extract and then analyze underlying themes from the chosen surahs. It basically involves text filtering, preprocessing, and tokenization before the application of the LDA algorithm. The identified topics were further validated by the Quranic scholars in order to validate their accuracy and theological consistency. Results show the effectiveness of topic modeling in religious text analysis, providing new insights into Quranic themes. This research not only furthers our understanding of the selected surahs but also provides a framework for applying computational techniques to religious text analysis, bridging traditional Islamic studies with modern data science approaches.

INTRODUCTION

Analyzing religious texts presents unique challenges due to their rich and complex nature [1]. Religious texts, such as the Holy Quran, are imbued with profound spiritual, moral, and ethical teachings that have shaped the lives of billions over centuries. These texts often contain multifaceted narratives, allegories, and instructions interwoven with linguistic subtleties and historical context [2]. The Quran, for instance, is written in Classical Arabic, a language that is highly nuanced and sophisticated [3]. This complexity is further compounded by the Quran's poetic structure, use of metaphors, and layers of meaning that can be interpreted in various ways.

One significant challenge in analyzing the Quran is preserving the context and meaning of the verses while performing computational analysis [4], [5]. The linguistic richness and depth of the Quran's content mean that any analytical approach must be capable of handling intricate semantic relationships and cultural references. Traditional analysis methods, such as manual exegesis (Tafsir), rely heavily on scholars' deep knowledge and understanding of the text, historical context, and linguistic nuances [6]. However, these methods can be subjective and time-consuming, limiting their scalability to large-scale text analysis [7].

Topic modeling [8], a well-known natural language processing technique, offers a promising computational approach to uncovering latent topics or subjects in large volumes of textual data [9]. It has applications in information retrieval, text mining [10], and computational social science [11]. Topic modeling can aid in developing subject summaries [12], enhancing information retrieval, and identifying evolving themes over time [13]. It is beneficial in analyzing social media data, predicting trends, opinions, and attitudes, and can be employed for document classification, recommendation systems, and content creation. As the volume of digital text data expands, topic modeling becomes an increasingly vital tool for researchers and businesses.

Latent Dirichlet Allocation (LDA) [14] is the most widely used algorithm for topic modeling. LDA is a probabilistic generative model that represents each document as a mixture of topics, with each topic being a probability distribution over words. Despite its scalability and wide range of applications, LDA does not account for correlations between topics, which can limit its effectiveness in specific contexts. The Correlated Topic Model (CTM) [15], which extends LDA by modeling correlations between topics, is more suitable for applications where

multiple issues are likely to co-occur, such as analyzing social media posts discussing various aspects of a single event [16].

In religious texts like the Holy Quran, topic analysis can provide valuable insights into the themes and subjects discussed within its verses. Sentiment analysis, an essential task in natural language processing, involves extracting and recognizing subjective information from text [17]. When applied to the Quran, sentiment analysis can reveal the emotional content and tone of the verses. However, traditional sentiment analysis methods might not adequately capture the Quran's rich and varied linguistic patterns. By finding the underlying themes [18] and subjects in the Quran and their link to sentiment, topic modeling can provide a supplementary method to sentiment analysis. Researchers can gain a deeper understanding of the emotional content of the Quran and its relevance to contemporary issues by combining topic modeling with sentiment analysis. This approach can enhance the comprehension of the Quran's influence on the beliefs and attitudes of Muslims worldwide. The complexity of the Quran, especially for non-Arabic-speaking Muslims, necessitates such advanced methodologies for meaningful topic extraction and analysis [19], [20].

In this study, we applied a comprehensive methodology for analyzing the topics present in the Holy Quran, explicitly focusing on Surah Al-Kahf and Surah An-Naml. We obtained a digital corpus of the Quranic text, which was carefully pre-processed to remove noise and ensure data quality. This involved removing stop words, punctuation, and numerical characters while preserving essential linguistic features. Next, we employed state-of-the-art topic modeling techniques, such as Latent Dirichlet Allocation (LDA), to extract latent topics from the Quranic verses. LDA enabled us to uncover hidden themes and subject areas within the Quran by leveraging the statistical relationships between words and topics. The extracted topics were then subjected to rigorous validation and interpretation by Islamic studies and computational linguistics experts. This methodology provided a systematic approach for topic extraction and ensured the results' accuracy and reliability.

Our contributions to this research are threefold. First, we unveiled the underlying thematic structure of the Holy Quran, highlighting recurring topics such as faith, morality, social justice, and spirituality. Second, our findings provided valuable insights into emphasizing specific subjects within the Quranic text. Finally, o ur methodology established a foundation for further research in comparative religious studies, Quranic exegesis, and the development of educational resources centered on specific Quranic topics. This study contributes to a deeper understanding of the Holy Quran and its teachings, facilitating scholarly exploration and fostering cross-disciplinary dialogue in religious studies.

OBJECTIVES

In this work, the modeling topic method LDA was utilized to be applied to the Quranic text in two different chapters, namely, Surah Al-Kahf and Surah An-Naml. The proposed method consists of several methods. In the first phase, we proposed collecting data from the digital corpus of the Holy Quran in text format. In the data pre-processing phase, the proposed method removes noise and irrelevant information (e.g., stop words, punctuation, numerical characters, etc.), tokenizing the text into individual words. Then, we proposed performing stemming to normalize words. The LDA algorithm was utilized for topic modeling by applying it to the pre-processed Quranic verses. LDA estimates the distribution of latent topics in the Quranic corpus. In the topic validation phase, we engage an Islamic studies and computational linguistics expert. Experts validate and interpret the extracted topics to ensure accuracy.

METHODS

The block diagram of the proposed topic modeling method of the Holy Quran using LDA is depicted in Fig. 1. In Fig. 1, the data collection phase of the proposed method obtains the digital corpus of the Holy Quran in text format. In the data pre-processing phase, the proposed method removes noise and irrelevant information (e.g., stop words, punctuation, numerical characters, etc.), tokenizing the text into individual words. Then, we proposed performing stemming to normalize words. The LDA algorithm was utilized for topic modeling by applying it to the pre-processed Quranic verses. LDA estimates the distribution of latent topics in the Quranic corpus. In the topic validation phase, we engage an Islamic studies and computational linguistics expert. Experts validate and interpret the extracted topics to ensure accuracy.

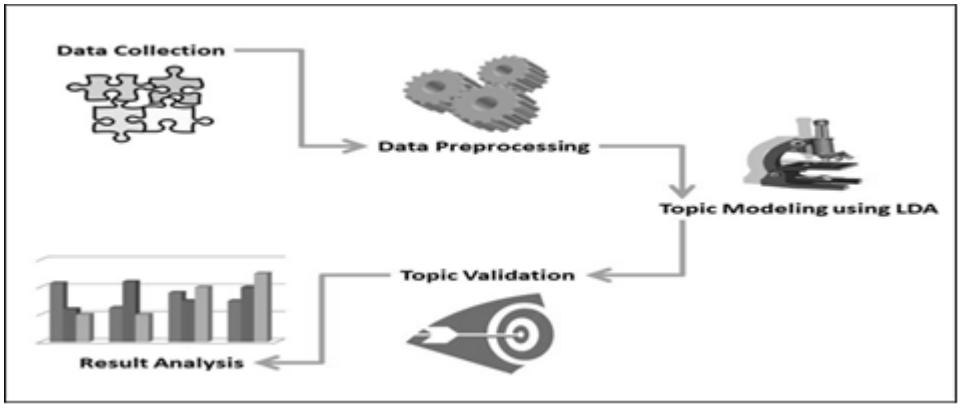

Figure 1 block diagram of the proposed work.

Data collection is the first step in this proposed work, which involves the collection of a digital corpus of the Holy Quran from reputable and reliable sources. This comprehensive dataset is the primary textual resource for the subsequent topic analysis. Care is taken to ensure that the collected dataset represents the entire Quranic text; special attention is given to preserving the original Arabic text, as it is the primary source for analysis. The quality and accuracy of the translations are carefully assessed to prevent any misinterpretation or loss of meaning during analysis. Through the data pre-processing phase, the collected Quranic text undergoes a thorough pre-processing process to ensure data quality and relevance. Noise and irrelevant elements, such as stop words, punctuation, and numerical characters, are removed. The text is then tokenized into individual words, and additional normalization techniques like stemming or lemmatization are applied to standardize words with common roots. This step prepares the dataset for subsequent topic modeling. Next, the LDA topic modeling was used to extract latent topics from the preprocessed Quranic verses. LDA is a powerful probabilistic generative model that estimates the distribution of topics within the Quranic corpus. By analyzing the statistical relationships between words and topics, LDA uncovers the hidden thematic structure of the Holy Quran, assigning each verse a probability distribution over different topics.Then, the phase of topic validation began to ensure the accuracy and credibility of the extracted topics, and some experts in Islamic studies and computational linguistics were engaged. These domain experts validate and interpret the results obtained from the LDA model. Their input is instrumental in verifying the identified topics' appropriateness and eliminating potential biases or errors introduced during the modeling process. In topic modeling using the Latent Dirichlet Allocation (LDA) algorithm, result validation is a critical step to ensure the quality and reliability of the outcomes, and this is the common validation technique used in topic modeling:Perplexity: commonly used metric to evaluate the coherence and generalization ability of the LDA model. A lower perplexity value indicates better model performance.Topic Coherence: Topic coherence measures the semantic coherence of the topics generated by the model. Higher coherence indicates more interpretable and meaningful topics.Visualization techniques: can help assess the interpretability and meaningfulness of the identified topics. Expert Evaluation: Engaging domain experts in the subject matter or textual data to evaluate the topics can provide valuable qualitative feedback on the relevance and correctness of the identified topics. Comparative Analysis: Comparing the results of different models, hyperparameter settings, or preprocessing techniques can help identify the best-performing configuration for the task.Hyperparameters are used to control various aspects while using the Latent Dirichlet Allocation (LDA) algorithm; hyperparameters are essential because they determine the model's behavior and can significantly impact its performance. The input to the LDA model is a document-term matrix (DTM) representing the corpus's frequency of words. In code, the hyperparameters used for topic modeling with the Latent Dirichlet Allocation (LDA) algorithm are as follows: Number of Topics (k): the number of topics (k) is not explicitly mentioned in the code, but it is specified during the LDA model creation using the LDA() function. The value of k represents the number of distinct topics the algorithm will try to identify in the text data. Alpha (α): the alpha hyperparameter, controlling the topic distribution for each document, is not explicitly set in the code. By default, the LDA() function in R uses an auto-tuning method to estimate alpha. The default behavior generally results in reasonable topic proportions for each document.

Eta (η): similarly, the eta hyperparameter, which controls the word distribution for each topic, is not explicitly set in the code. The LDA() function in R uses an auto-tuning method to estimate eta by default. This approach usually yields satisfactory word distributions for each topic. Seed: a random seed (seed value) of 1234 is used in the code to

ensure reproducibility. The seed is set when creating the LDA model, allowing the same results to be obtained in subsequent runs. As hyperparameters like Alpha and Eta are set using the auto-tuning methods provided by the LDA() function, the code aims to automatically find suitable values for these parameters. The Author did not specify fixed values for Alpha and Eta, relying on the defaults or auto-tuning. For this study, two prominent surahs from the Holy Quran, Surah Al-Kahf and Surah An-Naml, were selected as the primary dataset. Surah Al-Kahf, the 18th chapter of the Quran, holds significant importance in Islamic tradition, containing narratives of the "Companions of the Cave" and "Dhul-Qarnayn." Its themes of faith, righteousness, and divine guidance make it a source of inspiration and reflection for Muslims worldwide. On the other hand, Surah An-Naml, the 27th chapter, showcases the story of the Prophet Solomon, the ant, and the Queen of Sheba, highlighting the wisdom and magnificence granted by Allah. By including these two diverse surah in our dataset, I aimed to explore the thematic intricacies, textual patterns, and underlying wisdom present in the verses, contributing to a deeper understanding of the Quranic teachings and spiritual guidance.

RESULTS

The Quranic text for Surah Al-Kahf and Surah An-Naml is filtered from a larger dataset using the filter function. Relevant columns, including surah_id, ayah_id, ayah, surah_title_en, text, and ayah title, are selected to form the corpus. A list of stop words is loaded to aid in data preprocessing. This selection process is crucial for creating a focused corpus that includes only the necessary information for analysis. The Quranic text undergoes preprocessing to remove noise and irrelevant information. This includes eliminating punctuation and converting text to lowercase. Stop words are removed from the text using the list of loaded stop words. The preprocessed text is tokenized into individual words, a crucial step in preparing the data for topic modeling. Each document, representing a verse from Surah 18 and Surah 27, is tokenized into its constituent words. This step transforms the text data into a format suitable for further exploration. Text data from Surah An-Naml and Al-Kahf was preprocessed to facilitate further investigation. The text was initially filtered to include only Surah 18 and 27, and unnecessary punctuation was removed to focus on meaningful content. The resulting corpus contained 110, 93 documents, each representing a verse from Surah 18, 27. The documents were categorized into different groups based on the verse number, creating a structured dataset for analysis. As shown in Figs. 2 and 3, removing stop words is a critical step in enhancing the data quality. By eliminating common words that don't contribute much to the meaning, the analysis can focus on the substance of the text. This is particularly important when dealing with sacred texts like the Quran, where every word holds significance. This methodology is not just a technical analysis; it's a bridge between technology and religious studies. The careful creation of data and advanced analyses contribute to a scholarly exploration of sacred texts.

Figure 2 Surah Al-Kahf word cloud.

Figure 3 Surah An-Naml word cloud.

Keyword analysis was conducted to identify significant terms within the text. For example, the analysis revealed that the term "Allah" appeared prominently throughout the text, emphasizing its significance in Surah 18, 27. Additionally, the study highlighted phrases such as "Lord of the worlds," emphasizing the attributes of Allah. This keyword analysis provided insights into the recurring themes and key concepts in Surah 18, 27.

To gain a deeper understanding of the text, n-gram and compound analyses were performed. N-grams of various lengths were extracted to uncover patterns in the text. Also compound analysis was conducted to identify negated phrases, revealing instances where negation played a role in the text's meaning. These analyses helped uncover nuanced linguistic and semantic elements in Surah 18, 27, contributing to a richer interpretation of its content. Figs. 6 and 7.

Figure 6 Surah Al-Kahf elements.

Figure 7 Surah An-Naml elements.

Table 1: Surah Al-Kahf extracted topics

Topic by LDA algorithm	Validation by Expert	Starting Verse	Ending Verse
Story 1	Valid (1,31)	1	31
Parable 1	Valid (32,44)	32	44
Parable 2	Valid (45,59)	45	59
Story 2	Valid (60,82)	60	82
Story 3	Valid (83,102)	83	102

Extracting topics using Latent Dirichlet Allocation (LDA) from Surah Al-Kahf involves several steps. LDA is a generative probabilistic model that assumes documents are mixtures of topics, and each topic is a mixture of words. The success of LDA depends on the quality of preprocessing, parameter tuning, and the inherent structure of the text. It's both an art and a science to extract meaningful topics from a religious text.

Table 2: Surah An-Naml extracted topics

Topic by LDA algorithm	Validation by Expert	Starting Verse	Ending Verse
Parable 1	Valid (1,6)	1	6
Story 1	Valid (7,14)	7	14
Story 2	Valid (15,44)	15	44
Story 3	Valid (45,58)	45	58
Parable 2	Valid (59,93)	59	93

In surah, An-Naml, the LDA Algorithm showed three main stories: the first story is about prophet Musa, the second story is of Solomon, prophet and queen Bilquees, and the third story is of the people of Thamud and Saleh prophet and also two parables that came before and after the stories.

DISCUSSION

Latent Dirichlet Allocation (LDA) [14] and other topic modeling techniques have been widely applied across various domains in natural language processing (NLP), demonstrating their utility in tasks such as text classification, sentiment analysis, and information retrieval. In text classification, LDA has effectively identified underlying topics in a corpus and categorized documents accordingly. For example, Blei et al. (2003) [21] demonstrated the application of LDA for categorizing scientific articles based on discovered topics, showcasing its ability to handle large volumes of text and uncover hidden thematic structures. Sentiment analysis, another critical area in NLP, benefits from LDA by identifying the emotional content within texts, allowing researchers to discern patterns in attitudes and beliefs across different groups. LDA's flexibility and efficiency in handling extensive datasets make it a preferred choice for various text analysis applications.

However, LDA is not without limitations. The model assumes that words in a document are generated independently, which may not be realistic for all datasets. This assumption can lead to inaccuracies in capturing the nuanced relationships between words. Moreover, LDA's sensitivity to parameter choices, such as the number of topics, can significantly impact the quality of the results. Researchers have addressed some of these limitations through alternative approaches like the Correlated Topic Model (CTM). CTM extends LDA by allowing for topic correlations, providing a more nuanced understanding of topic relationships. For instance, Lu et al. (2018) [22] used CTM to analyze social media data on mental health, revealing that stress, anxiety, and depression frequently co-occur, highlighting the method's ability to uncover complex topic interdependencies. Similarly, Nguyen et al. (2018) [23] applied CTM to customer reviews, identifying correlations between usability and customer satisfaction, thus offering more profound insights into consumer preferences.

Beyond LDA and CTM, unigram-based methods have also proven effective in topic analysis. These methods focus on individual words' frequencies to identify significant topics and patterns. Liao et al. (2018) [24] used unigram-

based methods to analyze academic publications on big data, identifying key themes such as data analysis and machine learning. This approach's simplicity and effectiveness make it suitable for various applications. However, more complex models like LDA and CTM may not capture the same depth of relationships between topics. Jabeen et al. (2020) [25] combined unigrams with sentiment analysis to examine tweets on gender issues, revealing predominant topics and their associated sentiments. This study highlighted the public's concerns about gender equality and violence, demonstrating the potential of unigram-based methods for social media analysis.

Despite the advancements and applications of these methods, limited research focuses on applying topic modeling to religious texts, particularly the Holy Quran. Analyzing religious texts poses unique challenges due to their linguistic complexity, cultural significance, and the need to maintain interpretive integrity. Previous studies have focused on general text corpora or specific domains like social media and customer reviews. There is a clear gap in applying these advanced NLP techniques to religious texts rich in meaning and historical context.

This study aims to bridge this gap by employing LDA to analyze the Holy Quran, explicitly focusing on Surah Al-Kahf and Surah An-Naml. By extracting latent topics from these chapters, we seek to uncover the underlying thematic structure and provide deeper insights into the Quran's teachings. This work contributes to the growing field of digital humanities and offers a novel approach to understanding religious texts through computational methods. Our study stands on the shoulders of previous research in topic modeling while addressing the unique challenges posed by the intricate and layered nature of the Quranic text. Through rigorous validation and expert interpretation, we aim to ensure the accuracy and relevance of our findings, thereby advancing the field of Quranic studies and natural language processing.

CONCLUSIONS

In this work, the modeling topic method LDA was utilized to be applied on Quranic text on two different chapters, namely, Surah Al-Kahf and Surah An-Naml. The proposed method consists of several methods. In the first phase, we proposed collecting data from digital corpus of the Holy Quran in text format. In the data pre-processing phase, the proposed method removes noise and irrelevant information (e.g., stop words, punctuation, numerical characters, etc.), tokenizing the text into individual words. The, we proposed performing stemming to normalize words. For the sake of topic modelling, the LDA algorithm was utilized by applying it on the pre-processed Quranic verses. LDA estimates the distribution of latent topics in the Quranic corpus. In topic validation phase, we engage an expert in Islamic studies and computational linguistics. Experts validate and interpret the extracted topics to ensure accuracy.

REFRENCES

[1] Ali, Maulana Muhammad. Holy Quran. Ahmadiyya Anjuman Ishaat Islam Lahore USA, 2011.
[2] Al-Tarawneh, A. (2021). The role of Quran translations in radicalizing Muslims in the West and misrepresenting Islam. Journal of Religion and Violence, 9(1), 101-122.
[3] Faris, S., 2023. Exploring The Divine Message: Quranic Studies in The Context of Islamic Scholarship. Dirasah International Journal of Islamic Studies, 1(2), pp.111-125.
[4] Chukhanov, Sansyzbay, and Nurlan Kairb. "The importance of a semantic approach in understanding the texts of the Holy Quran and Sunnah." Pharos Journal of Theology 105, no. 3 (2024): 1-11.
[5] Taufik, Kurniawan, R., Ibrahim, R., Abdullah, H. and Widhiastuti, H., 2024. Preserving Qur'an Through Blind Eyes: Self-Regulation of Blind People in Memorizing the Qur'an. Journal of Disability & Religion, 28(1), pp.1-12.
[6] Shohoud, Yasser, Maged Shoman, and Sarah Abdelazim. "Quranic Conversations: Developing a Semantic Search tool for the Quran using Arabic NLP Techniques." arXiv preprint arXiv:2311.05120 (2023).
[7] Sutiyo, Febrian Rizki Adi, Nazruddin Safaat Harahap, Surya Agustian, and Reski Mai Candra. "Implementasi Question Answering Berbasis Chatbot Telegram Pada Tafsir Al-Jalalain Menggunakan Langchain dan LLM." KLIK: Kajian Ilmiah Informatika dan Komputer 4, no. 5 (2024): 2464-2472.
[8] Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W. and Hassan, A., 2023. Topic modeling algorithms and applications: A survey. Information Systems, 112, p.102131.
[9] Yu, D. and Xiang, B., 2023. Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling. Expert systems with applications, 225, p.120114.

[10] Wu, Zezhou, Qiufeng He, Jiarun Li, Guoqiang Bi, and Maxwell Fordjour Antwi-Afari. "Public attitudes and sentiments towards new energy vehicles in China: A text mining approach." Renewable and Sustainable Energy Reviews 178 (2023): 113242.
[11] Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. "Can large language models transform computational social science?." Computational Linguistics 50, no. 1 (2024): 237-291.
[12] Monir, E., & Salah, A. (2024). AraTSum: Arabic Twitter Trend Summarization Using Topic Analysis and Extractive Algorithms. International Journal of Computational Intelligence Systems, 17(1), 227.
[13] Yu, D. and Xiang, B., 2023. Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling. Expert systems with applications, 225, p.120114.
[14] Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey." Multimedia tools and applications 78 (2019): 15169-15211.
[15] Blei, David, and John Lafferty. "Correlated topic models." Advances in neural information processing systems 18 (2006): 147.
[16] Yalçın, Merve, Semanur Gürsoy, and Özcan Özyurt. "Mapping the Science World with Correlated Topic Modeling Analysis from Science-Related Posts on the Reddit Platform." Düzce Üniversitesi Bilim ve Teknoloji Dergisi 12, no. 3 (2024): 1664-1674.
[17] Alshammeri, Menwa, Eric Atwell, and Mhd Ammar Alsalka. "Quranic topic modelling using paragraph vectors." In Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 2, pp. 218-230. Springer International Publishing, 2021.
[18] Ta'a, Azman, Syuhada Zainal Abidin, Mohd Syazwan Abdullah, Abdul Bashah Mat Ali, and Muhammad Ahmad. "Al-Quran themes classification using ontology." (2012): 383-389.
[19] Bsoul, Qusay, Rosalina Abdul Salam, Jaffar Atwan, and Malik Jawarneh. "Arabic text clustering methods and suggested solutions for theme-based quran clustering: analysis of literature." Journal of Information Science Theory and Practice 9, no. 4 (2021): 15-34.
[20] Putra, S.J., Mantoro, T. and Gunawan, M.N., 2017, November. Text mining for Indonesian translation of the Quran: A systematic review. In 2017 International Conference on Computing, Engineering, and Design (ICCED) (pp. 1-5). IEEE.
[21] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.
[22] Lin, Kebin, et al. "Perovskite light-emitting diodes with external quantum efficiency exceeding 20 per cent." Nature 562.7726 (2018): 245-248.
[23] Nguyen, Nga Thi Thuy, et al. "RSAT 2018: regulatory sequence analysis tools 20th anniversary." Nucleic acids research 46.W1 (2018): W209-W214.
[24] Liao, Sheng-Kai, et al. "Satellite-relayed intercontinental quantum network." Physical review letters 120.3 (2018): 030501.
[25] Jabeen, Shahida, Xiaoying Gao, and Peter Andreae. "Semantic association computation: a comprehensive survey." Artificial Intelligence Review 53.6 (2020): 3849-3899.